DAHLGREN DIVISION
NAVAL SURFACE WARFARE CENTER
Dahlgren, Virginia 22448-5100

[4

NSWCDD/TR-95/92

DERIVATION OF JOINT REPRESENTATION MIXTURE MODEL EQUATIONS

BY GEORGE W. ROGERS AND RICHARD A. LOREY
SYSTEMS RESEARCH AND TECHNOLOGY DEPARTMENT

MAY 1995

Approved for public release; distribution is unlimited.

JGK

REPORT DOCUMENTATION PAGE

Form Approved
OBM No. 0704-0188

1. AGENCY USE ONLY (*Leave blank*)

2. REPORT DATE
May 1995

3. REPORT TYPE AND DATES COVERED

4. TITLE AND SUBTITLE
Derivation of Joint Representation Mixture Model Equations

5. FUNDING NUMBERS

6. AUTHOR(s)
George W. Rogers and Richard A. Lorey

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Surface Warfare Center
Dahlgren Division (Code B10)
17320 Dahlgren Rd.
Dahlgren, VA 22448-5100

8. PERFORMING ORGANIZATION REPORT NUMBER
NSWCDD/TR-95/92

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (*Maximum 200 words*)

One of the problems that arises in many large-scale applications of mixture models to density estimation is that, as the size of the data set increases, the class labeled data becomes a (proper) subset of the total data set. That is, while many small data sets may have all the observations labeled as to class membership, large data sets often consist of labeled subsets plus a potentially large unlabeled subset. Thus, it is desirable to have a unified framework for handling this combined supervised (class labeled data)/unsupervised (unlabeled data) problem. This is the motivation behind the following development of joint representation mixture models. The joint representation mixture model is defined, likelihood functions corresponding to different levels of data categorization with respect to class are presented, and the resultant iterative Expectation-Maximization equations are derived.

14. SUBJECT TERMS
Density Estimation, Mixture Models

15. NUMBER OF PAGES
23

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

FOREWORD

Density estimation plays a central role in probabilistic pattern recognition and signal processing. As data sets get larger, the cost of identifying a definitive class with each observation can become prohibitive. Instead, it becomes important to develop ways to process the data in ways that make use of all available information.

This work was done under the joint support of the Naval Surface Warfare Center, Dahlgren Division, Dahlgren, Virginia, Independent Research Program and the Office of Naval Research (R&T No. 4424314).

Approved by:

D. B. Colby

D. B. COLBY, Head
Systems Research and Technology Department

INTRODUCTION

Finite mixture models have proven to be quite flexible as parametric probability density function estimators.[1,2] Recently an adaptive mixture model was presented whose complexity or number of terms is determined in a data driven manner.[3] This approach has made possible the use of mixture models within a semi-parametric setting, and thus of much more general applicability/ utility than was possible under rigid parametric assumptions.

This semiparametric use of mixture models has resulted in efforts to develop alternative adaptive mixture model algorithms.[4,5] Recent applications of semiparametric mixture model density estimation can be found in References 6,7,8,9,10, and 11. Thus, in addition to the traditional parametric uses of mixture models, the semi-parametric application of mixture models is now well established.

One of the problems that arises in many large scale applications of mixture models to density estimation is that, as the size of the data set increases, the class labeled data becomes a subset of the total data set. That is, while many small data sets may have all the observations labeled as to class membership, large data sets often consist of labeled subsets plus a potentially large unlabeled subset.

The reason for this can be illustrated with an image processing example. Suppose that features are to be computed for each pixel for a number of images and that densities are to be computed for each class. Depending on the problem, the classes may correspond to vehicles, buildings, woods, and open terrain, or to tumorous and nontumorous tissue. If all the available data is to be used, the work in allocating each original pixel to one of the classes can easily become prohibitive. The more usual case is that only a representative subset of training data are class labeled with the balance either uncategorized or partially categorized. An example of the latter case is that it may be easy to say that there are no vehicles in this image, no buildings in that one, and so on but difficult or time consuming to identify each pixel corresponding to each class in each image. It is often the case in medical imagery that ground truth cannot be established definitively without a biopsy, again leading to less than full categorization of the observations.

Thus it is desirable to have a unified framework for handling this combined supervised (class labeled data)/unsupervised (unlabeled data) problem. This is the motivation behind the following development of joint representation mixture models.

This report begins with the formulation of traditional finite mixture models and proceeds to develop the joint representation mixture model equations.

JOINT REPRESENTATION MIXTURE MODELS

FINITE MIXTURE MODELS

Given a probability density function that can be represented as a finite (g term) mixture model

$$p(x|\psi) = \sum_{i=1}^{g} \pi_i f(x;\theta_i) , \tag{1}$$

where $f(\bullet;\theta)$ denotes a generic member of the chosen parametric family, the likelihood function for a set of n observations drawn from the particular density $p(x|\psi)$ can be written as

$$L(\psi) = \prod_{j=1}^{n} p(x_j|\psi) = \prod_{j=1}^{n} \sum_{i=1}^{g} \pi_i f(x_j;\theta_i) . \tag{2}$$

The vector θ_i represents the parameter set for the ith mixture component, while ψ represents the combined total parameter set including the mixing coefficients π_i. The log-likelihood function is

$$lnL(\psi) = \sum_{j=1}^{n} ln\left[\sum_{i=1}^{g} \pi_i f(x_j;\theta_i)\right]. \tag{3}$$

The maximum likelihood update equations can be obtained by taking derivatives of the log-likelihood function with respect to the mixture model parameters, setting the resulting expressions equal to zero, and solving for the parameters.

JOINT REPRESENTATION MIXTURE MODELS

The Joint Representation Mixture Model is defined by

$$p(x|\psi) = \sum_{i=1}^{g} p(\text{term } i)\, p(x|\text{term } i) \sum_{m=1}^{M} p(\text{class } m|\text{term } i) = \sum_{i=1}^{g} \pi_i f(x;\theta_i) \sum_{m=1}^{M} \zeta_{im} \tag{4}$$

where $\zeta_{im} = p(\text{class } m|\text{term } i)$ is an intra-term class mixing coefficient that gives the relative proportion of the ith term associated with the mth class with the constraint

$$\sum_{m=1}^{M} \zeta_{im} = \sum_{m=1}^{M} p\,(\text{class } m|\text{term } i) = 1\,. \tag{5}$$

This constraint merely says that for each term independently, the class mixing coefficients must sum to one, or equivalently, that an observation from term i must belong to one of the M classes with probability one.

The mixture model defined in Equations (4) and (5) represents a significant departure from traditional mixture model usage. Historically, a single mixture model has been used for either performing unsupervised clustering or to generate a probability density function for observations from a single class. If observations from multiple classes are to be dealt with, then a separate mixture model is developed for data from each class. This latter approach leaves open the question of how to incorporate partially class categorized or uncategorized observations when there are separate mixture models for each class. As will be seen, this new formulation leads to a unified treatment of these cases.

JOINT REPRESENTATION MIXTURE MODEL LIKELIHOOD FUNCTIONS

The likelihood function for class categorized data is

$$\begin{aligned} L_c(\psi) &= \prod_{j=1}^{n_c} \prod_{m=1}^{M} \{ [p\,(x_j|\,(\psi \cap \text{Class } m))\,]^{z_{jm}} \} \\ &= \prod_{j=1}^{n_c} \prod_{m=1}^{M} \left\{ \left[\sum_{i=1}^{g} \zeta_{im} \pi_i f(x_j;\theta_i) \right]^{z_{jm}} \right\}. \end{aligned} \tag{6}$$

Here, z_{jm} is a binary valued class indicator function. For observation j from class h, $z_{jh} = 1$ and $z_{jm} = 0$ for $m \neq h$. It thus can be considered as a picking function. It is used to "pick out" the desired contribution to the likelihood function. For each term in the product where it has the value zero, the contribution to the product is one so that the likelihood is unaffected. The log-likelihood function for class categorized data can be written

$$\begin{aligned} lnL_c(\psi) &= \sum_{j=1}^{n_c} \sum_{m=1}^{M} ln \left\{ \left[\sum_{i=1}^{g} \zeta_{im} \pi_i f(x_j;\theta_i) \right]^{z_{jm}} \right\} \\ &= \sum_{j=1}^{n_c} \sum_{m=1}^{M} z_{jm} ln \left\{ \sum_{i=1}^{g} \zeta_{im} \pi_i f(x_j;\theta_i) \right\} \\ &= \sum_{j=1}^{n_c} ln \sum_{i=1}^{g} \sum_{m=1}^{M} z_{jm} \zeta_{im} [\pi_i f(x_j;\theta_i)]\,. \end{aligned} \tag{7}$$

To derive the likelihood for partially (class) categorized data, first consider the likelihood appropriate for the case where the data is both class and term categorized. In this case the likelihood is

$$L_{ct}(\psi) = \prod_{j=1}^{n_c} \prod_{m=1}^{M} \prod_{i=1}^{g} \{ [\zeta_{im}\pi_i f(x_j;\theta_i)]^{z_{jm}z_{ji}} \} . \tag{8}$$

Where, as before, z_{jm} (z_{ji}) is a binary valued class (term) indicator function. For observation j from class h, $z_{jh} = 1$ and $z_{jm} = 0$ for $m \neq h$, while for observation j from term k, $z_{jk} = 1$ and $z_{ji} = 0$ for $i \neq k$.

In the absence of complete knowledge of z_{jm} and/or z_{ji} the usual procedure is to use expected values for either/both as

$$E(z_{jm}) = \xi_{jm} \tag{9}$$

and

$$E(z_{ji}) = \tau_{ij}^{p}. \tag{10}$$

Then for partially (class) categorized data, the likelihood is

$$L_p(\psi) = \prod_{j=1}^{n_p} \prod_{i=1}^{g} \prod_{m=1}^{M} \{ [\zeta_{im}\pi_i f(x_j;\theta_i)]^{\xi_{jm}\tau_{ij}^{p}} \} \tag{11}$$

whence

$$\begin{aligned} lnL_p(\psi) &= \sum_{j=1}^{n_p} \sum_{i=1}^{g} \sum_{m=1}^{M} \{ \xi_{jm}\tau_{ij}^{p} ln [\zeta_{im}\pi_i f(x_j;\theta_i)] \} \\ &= \sum_{j=1}^{n_p} \sum_{i=1}^{g} \sum_{m=1}^{M} ln \{ [\zeta_{im}\pi_i f(x_j;\theta_i)]^{\xi_{jm}\tau_{ij}^{p}} \} \end{aligned} \tag{12}$$

where ξ_{jm} is a prior or expected probability of class membership with

$$\sum_{m=1}^{M} [\xi_{jm}] = 1, \tag{13}$$

and

$$\tau_{ij}^{p} = \frac{\pi_i f(x_j;\theta_i)}{\sum_{i=1}^{g} \pi_i f(x_j;\theta_i)} \tag{14}$$

is the expectation (posterior probability) that the jth observation came from the ith mixture term. Since this is an expectation, it will be held fixed while taking derivatives. Two common methods for specifying partial categorization are (1) to give prior probabilities for each class for observation j or (2) to specify that the priors are zero for some subset of the classes and to use posterior probabilities across the remaining possible classes for the "unknown" ξ_{jm}.

For uncategorized data,

$$L_u(\psi) = \prod_{j=1}^{n_u}\left[\sum_{i=1}^{g}\sum_{m=1}^{M}\zeta_{im}\pi_i f(x_j;\theta_i)\right] = \prod_{j=1}^{n_u}\sum_{i=1}^{g}[\pi_i f(x_j;\theta_i)]\,, \tag{15}$$

$$lnL_u(\psi) = \sum_{j=1}^{n_u} ln\sum_{i=1}^{g}\sum_{m=1}^{M}\zeta_{im}\pi_i f(x_j;\theta_i) = \sum_{j=1}^{n_u} ln\sum_{i=1}^{g}[\pi_i f(x_j;\theta_i)]\,. \tag{16}$$

For combined categorized/partially categorized/uncategorized data,

$$lnL(\psi) = lnL_c(\psi) + lnL_p(\psi) + lnL_u(\psi) \tag{17}$$

or

$$\begin{aligned} lnL(\psi) &= \sum_{j=1}^{n_c} ln\sum_{i=1}^{g}\sum_{m=1}^{M} z_{jm}\zeta_{im}[\pi_i f(x_j;\theta_i)] \\ &+ \sum_{j=1}^{n_p}\sum_{m=1}^{M}\sum_{i=1}^{g}\xi_{jm}\tau_{ij} ln[\zeta_{im}\pi_i f(x_j;\theta_i)] + \sum_{j=1}^{n_u} ln\sum_{i=1}^{g}\pi_i f(x_j;\theta_i)\,. \end{aligned} \tag{18}$$

Historically with mixture models, reference to categorization of data has been with respect to which term of the mixture model is associated with a given observation. While this is logical when each term is ascribed a "class" status as in clustering, in this work a completely different definition of categorized data is being dealt with. In this case, the concern is that of categorizing data only with respect to class membership, rather than with respect to individual mixture model terms.

DERIVATION OF MAXIMUM LIKELIHOOD E-M EQUATIONS

The log-likelihood is to be maximized with respect to variation of the parameter set ψ. The usual procedure of taking the derivative of the log-likelihood with respect to each of the parameters, setting the resultant expressions equal to zero, and solving for resultant expressions that must be satisfied for a maxima is followed. Since these expressions comprise a coupled set of nonlinear equations, convergence must be obtained through iteration which is just the E-M algorithm.[1]

While the resultant equations for ζ and π are independent of the parametric function $f(\bullet;\theta)$ used for the mixture model, the remaining equations are not. Therefore, in the following derivation, it is assumed that

$$f(x;\theta_i) = N(x;\mu_i,\Sigma_i) = \frac{1}{\sqrt{2\pi\Sigma_i}}\exp\left[-\frac{(x-\mu_i)^2}{2\Sigma_i}\right]. \tag{19}$$

That is, the mixture models are mixtures of univariate normals, each of which is parametrized by mean μ and variance Σ.

E-M Equations for ζ

First, derive the equations for the parameter ζ. Taking the derivatives of each of the three log-likelihood components results in the following expressions.

$$\frac{\partial}{\partial\zeta_{im}} lnL_c(\psi) = \sum_{j=1}^{n_c} \frac{\left(z_{jm}\zeta_{im}\zeta_{im}^{-1} - z_{jM}\zeta_{iM}\zeta_{iM}^{-1}\right)\pi_i f(x_j;\theta_i)}{\sum_{i=1}^{g}\sum_{m=1}^{M} z_{jm}\zeta_{im}\pi_i f(x_j;\theta_i)}, \tag{20}$$

$$\frac{\partial}{\partial\zeta_{im}} lnL_p(\psi) = \sum_{j=1}^{n_p}\left(\xi_{jm}\zeta_{im}^{-1} - \xi_{jM}\zeta_{iM}^{-1}\right)\tau_{ij}^{p}, \tag{21}$$

$$\frac{\partial}{\partial\zeta_{im}} lnL_u(\psi) = 0. \tag{22}$$

Combining these expressions and setting equal to zero,

$$\frac{\partial}{\partial\zeta_{im}}[lnL_c(\psi) + lnL_p(\psi) + lnL_u(\psi)] = 0, \tag{23}$$

gives

$$\sum_{j=1}^{n_c} \frac{z_{jm}\zeta_{im}\pi_i f(x_j;\theta_i)}{\sum_{i=1}^{g}\sum_{m=1}^{M} z_{jm}\zeta_{im}\pi_i f(x_j;\theta_i)} + \sum_{j=1}^{n_p}\xi_{jm}\tau_{ij}^{p} = C_i\zeta_{im}, \tag{24}$$

where the constant arises by virtue of the constraint among the ζ_{im} for fixed i. Equation (23) must hold independently for every i. Define τ_{ijm} by

$$\tau_{ijm} = \frac{z_{jm}\zeta_{im}\pi_i f(x_j;\theta_i)}{\sum_{i=1}^{g}\sum_{m=1}^{M} z_{jm}\zeta_{im}\pi_i f(x_j;\theta_i)}, \tag{25}$$

and similarly define both τ_{ij}^c and ζ_{im_j} through

$$\tau_{ij}^c = \sum_{m=1}^{M}\tau_{ijm} = \frac{\zeta_{im_j}\pi_i f(x_j;\theta_i)}{\sum_{i=1}^{g}\zeta_{im_j}\pi_i f(x_j;\theta_i)} = \frac{\sum_{m=1}^{M} z_{jm}\zeta_{im}\pi_i f(x_j;\theta_i)}{\sum_{i=1}^{g}\sum_{m=1}^{M} z_{jm}\zeta_{im}\pi_i f(x_j;\theta_i)}. \tag{26}$$

Then Equation (24) becomes

$$\sum_{j=1}^{n_c}\tau_{ijm} + \sum_{j=1}^{n_p}\xi_{jm}\tau_{ij}^p = C_i\zeta_{im}. \tag{27}$$

By summing both sides over m for fixed i, and using the identity Equation (5), the constant is found to be

$$C_i = \sum_{j=1}^{n_c}\tau_{ij}^c + \sum_{j=1}^{n_p}\tau_{ij}^p. \tag{28}$$

Thus

$$\zeta_{im} = \left[\frac{1}{\sum_{j=1}^{n_c}\tau_{ij}^c + \sum_{j=1}^{n_p}\tau_{ij}^p}\right]\left[\sum_{j=1}^{n_c}\tau_{ijm} + \sum_{j=1}^{n_p}\xi_{jm}\tau_{ij}^p\right], \tag{29}$$

which completes the derivation of the E-M equation for the ζ.

E-M Equations for π

Next, derive the corresponding equation for π by evaluating

$$\frac{\partial}{\partial \pi_i}[lnL_c(\psi) + lnL_p(\psi) + lnL_u(\psi)] = 0. \tag{30}$$

Recalling that

$$lnL(\psi) = \sum_{j=1}^{n_c} ln \sum_{i=1}^{g} \sum_{m=1}^{M} z_{jm}\zeta_{im}[\pi_i f(x_j;\theta_i)]$$
$$+ \sum_{j=1}^{n_p} \sum_{m=1}^{M} \sum_{i=1}^{g} \xi_{jm}\tau_{ij} ln[\zeta_{im}\pi_i f(x_j;\theta_i)] + \sum_{j=1}^{n_u} ln \sum_{i=1}^{g} \pi_i f(x_j;\theta_i),$$

it is found that

$$\frac{\partial}{\partial \pi_i} lnL(\psi) = \sum_{j=1}^{n_c} [T_c(j,i) - T_c(j,g)] + \sum_{j=1}^{n_p} [T_p(j,i) - T_p(j,g)]$$
$$+ \sum_{j=1}^{n_u} [T_u(j,i) - T_u(j,g)], \tag{31}$$

where

$$T_c(j,i) = \frac{\sum_{m=1}^{M} z_{jm}\zeta_{im}[f(x_j;\theta_i)]}{\sum_{i=1}^{g}\sum_{m=1}^{M} z_{jm}\zeta_{im}[\pi_i f(x_j;\theta_i)]} = \sum_{m=1}^{M} \frac{\tau_{ijm}}{\pi_i} = \frac{\tau_{ij}^c}{\pi_i}, \tag{32}$$

$$T_c(j,g) = -\sum_{m=1}^{M}\left[\frac{\tau_{gjm}}{\pi_g}\right] = -\frac{\tau_{gj}^c}{\pi_g}, \tag{33}$$

$$T_p(j,i) = \sum_{m=1}^{M} \frac{\{\xi_{jm}\tau_{ij}^p\}}{\pi_i} = \frac{\tau_{ij}^p}{\pi_i}, \tag{34}$$

and

$$T_u(j,i) = \frac{f(x_j;\theta_i)}{\sum_{i=1}^{g}[\pi_i f(x_j;\theta_i)]} = \frac{\tau_{ij}^u}{\pi_i}. \tag{35}$$

In terms of these expressions, the derivative is given by

$$\begin{aligned} &\frac{\partial}{\partial \pi_i} lnL(\psi) \\ &= \sum_{j=1}^{n_c}[T_c(j,i) - T_c(j,g)] + \sum_{j=1}^{n_p}[T_p(j,i) - T_p(j,g)] + \sum_{j=1}^{n_u}[T_u(j,i) - T_u(j,g)] \qquad (36) \\ &= \sum_{j=1}^{n_c}\left[\frac{\tau_{ij}^c}{\pi_i} - \frac{\tau_{gj}^c}{\pi_g}\right] + \sum_{j=1}^{n_p}\left[\frac{\tau_{ij}^p}{\pi_i} - \frac{\tau_{gj}^p}{\pi_g}\right] + \sum_{j=1}^{n_u}\left[\frac{\tau_{ij}^u}{\pi_i} - \frac{\tau_{gj}^u}{\pi_g}\right]. \end{aligned}$$

Setting this equal to zero results in

$$\sum_{j=1}^{n_c}\left[\frac{\tau_{ij}^c}{\pi_i}\right] + \sum_{j=1}^{n_p}\left[\frac{\tau_{ij}^p}{\pi_i}\right] + \sum_{j=1}^{n_u}\left[\frac{\tau_{ij}^u}{\pi_i}\right] = \sum_{j=1}^{n_c}\left[\frac{\tau_{gj}^c}{\pi_g}\right] + \sum_{j=1}^{n_p}\left[\frac{\tau_{gj}^p}{\pi_i}\right] + \sum_{j=1}^{n_u}\left[\frac{\tau_{gj}^u}{\pi_i}\right] = C. \tag{37}$$

Solving for C,

$$\pi_i C = \sum_{j=1}^{n_c}\tau_{ij}^c + \sum_{j=1}^{n_p}\tau_{ij}^p + \sum_{j=1}^{n_u}\tau_{ij}^u, \tag{38}$$

whence

$$C = \sum_{i=1}^{g}\pi_i C = \sum_{j=1}^{n_c}\sum_{i=1}^{g}\tau_{ij}^c + \sum_{j=1}^{n_p}\sum_{i=1}^{g}\tau_{ij}^p + \sum_{j=1}^{n_u}\sum_{i=1}^{g}\tau_{ij}^u = n_c + n_p + n_u. \tag{39}$$

The final result is thus

$$\pi_i = \frac{1}{(n_c + n_p + n_u)}\left[\sum_{j=1}^{n_c}\tau_{ij}^c + \sum_{j=1}^{n_p}\tau_{ij}^p + \sum_{j=1}^{n_u}\tau_{ij}^u\right]. \tag{40}$$

This gives the maximum likelihood equation for π.

E-M Equations for μ

Next, derive the corresponding equation for μ. Start with

$$\frac{\partial}{\partial\mu_i}[lnL_c(\psi) + lnL_p(\psi) + lnL_u(\psi)] = 0. \tag{41}$$

Recalling,

$$lnL(\psi) = \sum_{j=1}^{n_c} ln \sum_{i=1}^{g} \sum_{m=1}^{M} z_{jm}\zeta_{im}[\pi_i f(x_j;\theta_i)]$$
$$+ \sum_{j=1}^{n_p} \sum_{m=1}^{M} \sum_{i=1}^{g} \xi_{jm}\tau_{ij}^{p} ln[\zeta_{im}\pi_i f(x_j;\theta_i)] + \sum_{j=1}^{n_u} ln \sum_{i=1}^{g} \pi_i f(x_j;\theta_i),$$

gives

$$\frac{\partial}{\partial\mu_i} lnL(\psi) = \sum_{j=1}^{n_c}[T_c(j,i)] + \sum_{j=1}^{n_p}[T_p(j,i)] + \sum_{j=1}^{n_u}[T_u(j,i)]. \tag{42}$$

Where

$$T_c(j,i) = \frac{[\zeta_{im_j}\pi_i f(x_j;\theta_i)][(x_j-\mu_i)/\Sigma_{ii}]}{\sum_{i=1}^{g}\zeta_{im_j}\pi_i f(x_j;\theta_i)} = \sum_{m=1}^{M}\tau_{ijm}[(x_j-\mu_i)/\Sigma_{ii}] \tag{43}$$

for the univariate normal mixture model case. Similarly,

$$T_p(j,g) = \sum_{m=1}^{M}\{\xi_{jm}\tau_{ij}^{p}\}[(x_j-\mu_i)/\Sigma_{ii}] = \tau_{ij}^{p}[(x_j-\mu_i)/\Sigma_{ii}] \tag{44}$$

and

$$T_u(j,i) = \tau_{ij}^{u}[(x_j-\mu_i)/\Sigma_{ii}]. \tag{45}$$

When put together, the resulting expression is

$$\frac{\partial}{\partial\mu_i} lnL(\varphi) = \sum_{j=1}^{n_c} [T_c(j,i)] + \sum_{j=1}^{n_p} [T_p(j,i)] + \sum_{j=1}^{n_u} [T_u(j,i)]$$
$$= \sum_{j=1}^{n_c} \sum_{m=1}^{M} [\tau_{ijm}(x_j - \mu_i)/\Sigma_{ii}] + \sum_{j=1}^{n_p} \left[\tau_{ij}^{p}(x_j - \mu_i)/\Sigma_{ii}\right] + \sum_{j=1}^{n_u} \left[\tau_{ij}^{u}(x_j - \mu_i)/\Sigma_{ii}\right] = 0, \tag{46}$$

which becomes

$$\mu_i \left[\sum_{j=1}^{n_c} \left[\tau_{ij}^{c}\right] + \sum_{j=1}^{n_p} \left[\tau_{ij}^{p}\right] + \sum_{j=1}^{n_u} \left[\tau_{ij}^{u}\right]\right] = \sum_{j=1}^{n_c} \left[\tau_{ij}^{c} x_j\right] + \sum_{j=1}^{n_p} \left[\tau_{ij}^{p} x_j\right] + \sum_{j=1}^{n_u} \left[\tau_{ij}^{u} x_j\right], \tag{47}$$

whence the final result is obtained

$$\mu_i = \frac{\sum_{j=1}^{n_c} \left[\tau_{ij}^{c} x_j\right] + \sum_{j=1}^{n_p} \left[\tau_{ij}^{p} x_j\right] + \sum_{j=1}^{n_u} \left[\tau_{ij}^{u} x_j\right]}{\left[\sum_{j=1}^{n_c} \left[\tau_{ij}^{c}\right] + \sum_{j=1}^{n_p} \left[\tau_{ij}^{p}\right] + \sum_{j=1}^{n_u} \left[\tau_{ij}^{u}\right]\right]}. \tag{48}$$

This provides the joint representation mixture model expression for μ.

<u>E-M Equations for Σ</u>

Finally, derive the corresponding equation for Σ. Evaluating

$$\frac{\partial}{\partial\Sigma_{ii}} [lnL_c(\psi) + lnL_p(\psi) + lnL_u(\psi)] = 0, \tag{49}$$

with

$$lnL(\psi) = \sum_{j=1}^{n_c} ln \sum_{i=1}^{g} \sum_{m=1}^{M} z_{jm}\zeta_{im}[\pi_i f(x_j;\theta_i)]$$
$$+ \sum_{j=1}^{n_p} \sum_{m=1}^{M} \sum_{i=1}^{g} \xi_{jm}\tau_{ij} ln[\zeta_{im}\pi_i f(x_j;\theta_i)] + \sum_{j=1}^{n_u} ln \sum_{i=1}^{g} \pi_i f(x_j;\theta_i) .$$

results in

$$\frac{\partial}{\partial \Sigma_i} lnL(\psi) = \sum_{j=1}^{n_c} [T_c(j,i)] + \sum_{j=1}^{n_p} [T_p(j,i)] + \sum_{j=1}^{n_u} [T_u(j,i)] \tag{50}$$

where

$$T_c(j,i) = \sum_{m=1}^{M} \tau_{ijm} \left[\frac{(x_j - \mu_i)^2}{2\Sigma_{ii}^2} - \frac{1}{2\Sigma_{ii}} \right] \tag{51}$$

where again, attention has been restricted to the univariate normal mixture model case. Similarly,

$$T_p(j,g) = \tau_{ij}^p \left[\frac{(x_j - \mu_i)^2}{2\Sigma_{ii}^2} - \frac{1}{2\Sigma_{ii}} \right], \tag{52}$$

and

$$T_u(j,i) = \tau_{ij}^u \left[\frac{(x_j - \mu_i)^2}{2\Sigma_{ii}^2} - \frac{1}{2\Sigma_{ii}} \right]. \tag{53}$$

When put together, the resulting expression is

$$\begin{aligned}
\frac{\partial}{\partial \Sigma_i} lnL(\psi) &= \sum_{j=1}^{n_c} [T_c(j,i)] + \sum_{j=1}^{n_p} [T_p(j,i)] + \sum_{j=1}^{n_u} [T_u(j,i)] \\
&= \left\{ \sum_{j=1}^{n_c} \tau_{ij}^c \left[\frac{(x_j - \mu_i)^2}{2\Sigma_{ii}^2} - \frac{1}{2\Sigma_{ii}} \right] + \sum_{j=1}^{n_p} \tau_{ij}^p \left[\frac{(x_j - \mu_i)^2}{2\Sigma_{ii}^2} - \frac{1}{2\Sigma_{ii}} \right] + \sum_{j=1}^{n_u} \tau_{ij}^u \left[\frac{(x_j - \mu_i)^2}{2\Sigma_{ii}^2} - \frac{1}{2\Sigma_{ii}} \right] \right\} \\
&= 0,
\end{aligned} \tag{54}$$

which leads to the final result

$$\Sigma_i = \frac{\sum_{j=1}^{n_c} \tau_{ij}^c (x_j - \mu_i)^2 + \sum_{j=1}^{n_p} \tau_{ij}^p (x_j - \mu_i)^2 + \sum_{j=1}^{n_u} \tau_{ij}^u (x_j - \mu_i)^2}{\left[\sum_{j=1}^{n_c} \tau_{ij}^c + \sum_{j=1}^{n_p} \tau_{ij}^p + \sum_{j=1}^{n_u} \tau_{ij}^u \right]}. \tag{55}$$

This provides the joint representation mixture model expression for Σ.

RESULTS

SUMMARY OF UNIVARIATE ITERATIVE E-M UPDATE EQUATIONS

$$\zeta_{im} = \left[\frac{1}{\sum_{j=1}^{n_c} \tau_{ij}^{c} + \sum_{j=1}^{n_p} \tau_{ij}^{p}}\right]\left[\sum_{j=1}^{n_c} \tau_{ijm} + \sum_{j=1}^{n_p} \xi_{jm}\tau_{ij}^{p}\right], \tag{56}$$

$$\pi_i = \frac{1}{(n_c + n_p + n_u)}\left[\sum_{j=1}^{n_c} \tau_{ij}^{c} + \sum_{j=1}^{n_p} \tau_{ij}^{p} + \sum_{j=1}^{n_u} \tau_{ij}^{u}\right], \tag{57}$$

$$\mu_i = \frac{\sum_{j=1}^{n_c} \left[\tau_{ij}^{c} x_j\right] + \sum_{j=1}^{n_p} \left[\tau_{ij}^{p} x_j\right] + \sum_{j=1}^{n_u} \left[\tau_{ij}^{u} x_j\right]}{\left[\sum_{j=1}^{n_c} \left[\tau_{ij}^{c}\right] + \sum_{j=1}^{n_p} \left[\tau_{ij}^{p}\right] + \sum_{j=1}^{n_u} \left[\tau_{ij}^{u}\right]\right]}, \tag{58}$$

$$\Sigma_i = \frac{\sum_{j=1}^{n_c} \tau_{ij}^{c} (x_j - \mu_i)^2 + \sum_{j=1}^{n_p} \tau_{ij}^{p} (x_j - \mu_i)^2 + \sum_{j=1}^{n_u} \tau_{ij}^{u} (x_j - \mu_i)^2}{\left[\sum_{j=1}^{n_c} \tau_{ij}^{c} + \sum_{j=1}^{n_p} \tau_{ij}^{p} + \sum_{j=1}^{n_u} \tau_{ij}^{u}\right]}, \tag{59}$$

where

$$\tau_{ijm} = \frac{z_{jm}\zeta_{im}\pi_i f(x_j;\theta_i)}{\sum_{i=1}^{g} \sum_{m=1}^{M} z_{jm}\zeta_{im}\pi_i f(x_j;\theta_i)} \tag{60}$$

and

$$\tau_{ij}^{c} = \sum_{m=1}^{M} \tau_{ijm} = \frac{\zeta_{im_j} \pi_i f(x_j;\theta_i)}{\sum_{i=1}^{g} \zeta_{im_j} \pi_i f(x_j;\theta_i)} = \frac{\sum_{m=1}^{M} z_{jm} \zeta_{im} \pi_i f(x_j;\theta_i)}{\sum_{i=1}^{g} \sum_{m=1}^{M} z_{jm} \zeta_{im} \pi_i f(x_j;\theta_i)}. \tag{61}$$

Similarly

$$\tau_{ij}^{p} = \tau_{ij}^{u} = \frac{\pi_i f(x_j;\theta_i)}{\sum_{i=1}^{g} \pi_i f(x_j;\theta_i)} \tag{62}$$

where it is to be remembered that τ_{ij}^{p} is only computed for those x_j that are partially categorized observations and similarly for τ_{ij}^{u}.

The E-M algorithm then consists of iterating the Expectation step consisting of evaluating Equations (60), (61), and (62) for the appropriate observations and the Maximization step, which consists of evaluating new parameter values using Equations. (56) through (59).

The multivariate versions can be obtained by making x_j and μ_i vector quantities and Σ_i a matrix. The equations for μ_i and Σ_i become

$$\mu_i^k = \frac{\sum_{j=1}^{n_c} \left[\tau_{ij}^{c} x_j^k\right] + \sum_{j=1}^{n_p} \left[\tau_{ij}^{p} x_j^k\right] + \sum_{j=1}^{n_u} \left[\tau_{ij}^{u} x_j^k\right]}{\left[\sum_{j=1}^{n_c} \left[\tau_{ij}^{c}\right] + \sum_{j=1}^{n_p} \left[\tau_{ij}^{p}\right] + \sum_{j=1}^{n_u} \left[\tau_{ij}^{u}\right]\right]} \tag{63}$$

and

$$\Sigma_i^{kl} = \frac{\sum_{j=1}^{n_c} \tau_{ij}^{c}\left(x_j^k - \mu_i^k\right)\left(x_j^l - \mu_i^l\right) + \sum_{j=1}^{n_p} \tau_{ij}^{p}\left(x_j^k - \mu_i^k\right)\left(x_j^l - \mu_i^l\right) + \sum_{j=1}^{n_u} \tau_{ij}^{u}\left(x_j^k - \mu_i^k\right)\left(x_j^l - \mu_i^l\right)}{\left[\sum_{j=1}^{n_c} \tau_{ij}^{c} + \sum_{j=1}^{n_p} \tau_{ij}^{p} + \sum_{j=1}^{n_u} \tau_{ij}^{u}\right]} \tag{64}$$

where the component indices are denoted by superscripts.

Finally, once the joint representation mixture model has been obtained based on any combination of class categorized, partial class categorized, and uncategorized data, if desired the probability density function for an individual class can be obtained through

$$\hat{p}\,(x|\text{Class y}, \psi) = \frac{\sum_{i=1}^{g} \zeta_{iy}\pi_i f(x;\theta_i)}{\sum_{i=1}^{g} \zeta_{iy}\pi_i}. \tag{65}$$

This gives a properly normalized mixture model density estimate for an individual class.

CONCLUSIONS

The derivation of the joint representation mixture model E-M equations has been presented for mixtures of normal components. While the detailed derivation was for the univariate case, a slightly more complicated derivation results in the full multivariate equations. The results for this case have been presented without derivation. It is also important to note that the equations derived for ζ and π are not restricted to the use of normal functions in the mixture model. These equations are entirely general for the class of joint representation mixture models considered.

The joint representation approach represents a significant philosophical departure from current mixture model usage. The standard mixture model usage is either to build a separate mixture model for each class when the observations are class labeled or to assume that each class is normally distributed so that a mixture model for all the data can be interpreted as a mixture of normal classes. The approach, in effect, totally relaxes the requirement for each class to be normally distributed. Philosophically, a semiparametric viewpoint has been taken in that it is assumed that each class can be modelled by a (potentially complex) mixture model and that no significance is to be ascribed to an individual term in the mixture. As an example, contrast a mixture model approximation to a lognormal density to a mixture of two normals. In the latter case, it may well make sense to care about which of the two terms gave rise to a particular observation. That is, each term may correspond to some recognizable class. However, in the lognormal case where by assumption the density is nonparametric with respect to representation by normal mixtures (even though it has been modelled it with a parametric approximation), this sort of distinction has little or no meaning. In other words, none of the terms used to model the lognormal density can be given any independent or class meaning.

This approach is thus appropriate for combined supervised/unsupervised (various levels of class categorization) learning when the individual class densities may be more complex than simple normals. It provides a unified framework for handling this problem. Once the joint representation density has been estimated, densities corresponding to the individual classes can be easily recovered.

Future reports will detail the derivation of recursive versions of these equations as well as a method for determining the complexity of the joint representation mixture model in a data driven manner.

REFERENCES

1. Titterington, D. M.; Smith, A. F. M.; and Makov, U. E., *Statistical Analysis of Finite Mixture Distributions*, Wiley, New York, 1985.

2. McLachlan, G. J. and Basford, K. E., *Mixture Models*, Marcel Dekker, New York, 1988.

3. Priebe, C. E., "Adaptive Mixtures", *J. Am. Statist. Assoc.*, Vol. 89, pp. 796-806, 1994.

4. Rogers, G. W., To be Published 1995.

5. Solka, J. L., Matching Model Information Content to Data Information, Ph.D. Thesis, George Mason Univ., 1995.

6. Priebe, C. E.; Julin, E. G.; Rogers, G. W.; Healy, D. M.; Lu , J.; Solka, J. L.; and Marchette, D. J., "Incorporating Segmentation Boundaries into the Calculation of Fractal Dimension Features," *Proc. 26th Symp. on the Interface*, Research Triangle Park, NC, 15-18 Jun 1994, pp. 52-56.

7. Priebe, C. E.; Lorey, R. A.; Marchette, D. J.; Solka, J. L.; and Rogers, G. W., "Nonparametric spatio-temporal change point analysis for early detection in mammography," *Proc. of the Second Int'l Workshop on Digital Mammography (SIWDM)*, York, UK, 10-12 Jul 1994, pp. 111-120.

8. Priebe, C. E.; Rogers, G. W.; Marchette, D. J.; and Solka, J. L., "Change Point Analysis With Adaptive Mixture Models," *Proc. of the Int'l Geoscience and Remote Sensing Symp. (IGARSS '94)* ,Pasadena, CA, 8-12 Aug 1994, pp. 1795-1798.

9. Solka, J. L.; Priebe, C. E.; Rogers, G. W.; Poston, W. L.; and Lorey, R. A., "Maximum Likelihood Density Estimation with Term Creation and Annihilation," *Proc. 26th Symp. on the Interface*, Research Triangle Park, NC, 15-18 Jun 1994, pp. 222-225.

10. Priebe, C. E.; Solka, J. L.; Lorey, R. A.; Rogers, G. W.; Poston, W. L.; Kallergi, M.; Qian,W.; Clarke, L. P.; and Clark, R. A., "The Application of Fractal Analysis to Mammographic Tissue Classification," *Cancer Letters*, Vol. 77, 1994, pp. 183-189.

11. Solka, J. L.; Priebe, C. E.; and Rogers, G. W., "An Initial Assessment of Discriminant Surface Complexity for Power Law Features," *Simulation: Special Issue: Neural Networks: Model Development or Applications*, Vol. 58, 1992, pp. 311-318.

DISTRIBUTION

	COPIES
DOD ACTIVITIES (CONUS)	
ATTN CODE 342PS (HAWKINS)	1
OFFICE OF NAVAL RESEARCH	
800 N QUINCY ST	
ARLINGTON VA 22217	
ATTN CODE E29L	
(TECHNICAL LIBRARY)	1
COMMANDING OFFICER	
CSSDD NSWC	
6703 W HIGHWAY 98	
PANAMA CITY FL 32407-7001	
DEFENSE TECHNICAL INFORMATION CTR	
CAMERON STATION	
ALEXANDRIA VA 22304-6145	12
NON-DOD ACTIVITIES (CONUS)	
THE CNA CORPORATION	
PO BOX 16268	
ALEXANDRIA VA 22302-0268	1
ATTN GIFT AND EXCHANGE DIVISION	4
LIBRARY OF CONGRESS	
WASHINGTON DC 20540	
INTERNAL	
B10 (ROGERS)	10
B10 (LOREY)	10
B10 (MARCHETTE)	1
B10 (SOLKA)	1
C	1
D	1
D4	1
E231	3
E282 (SWANSBURG)	1
G33 (POSTON)	1
N74 (GIDEP)	1